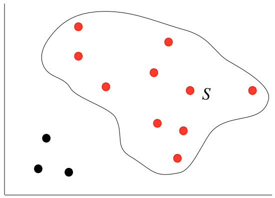
# Unsupervised Support Vector Machines

Manel Martínez-Ramón

ECE, UNM

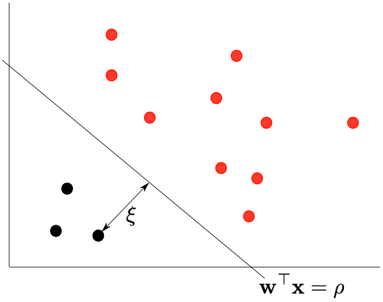
October, 2018

- Idea of SV Novelty Detection: Assume some dataset drawn from a latent probability distribution $P$. Estimate a simple subset $S$ of input space such that the probability that a test point drawn from $P$ lies outside of $S$ equals some a priori specified $\nu$ between 0 and 1. (Schölkopf et al 2001).




- The problem is classical, and it can be solved from the point of view of SLT by imposing positiveness in $S$ and negativeness in the complement while at the same time maximizing a margin.

- The strategy adopted assumes that almost all the data can be separated from the origin with a hyperplane, and only a small subset with probability $\nu$ will be in the space between the hyperplane and the origin. In order to confine the *normal data* in the smallest possible space, we maximize the distance of the hyperplane to the center.



- This is very restrictive, but the limitations disappear when the formulation is extended to kernel spaces.
- The formulation of the primal optimization is

$$\text{Minimize } ||\mathbf{w}||^2 + \frac{1}{N\nu}\sum_{n=1}^{N}\xi_n - \rho$$

$$\text{subject to } \begin{cases} \mathbf{w}^\top \mathbf{x}_n \geq \rho - \xi_n \\ \xi_n \geq 0 \end{cases}$$

- The corresponding dual is

$$\text{Minimize } \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$
$$\text{subject to } \begin{cases} 0 \leq \alpha_n \leq \frac{1}{N\nu} \\ \sum_{n=1} \alpha_i = 1 \end{cases}$$

- This minimization can be solved by QP.
- Since any $\mathbf{x_n}$ for which $0 < \alpha_n < \frac{1}{N\nu}$ satisfies the equality $\mathbf{w}^\top \mathbf{x}_n + \rho = 0$, $\rho$ can be easily recovered.

Assume the solution of

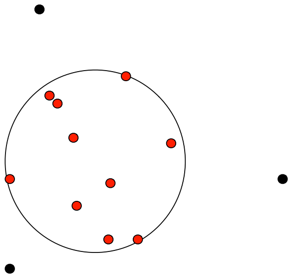
$$\mathbf{w}^\top \mathbf{x}_n \geq \rho - \xi_n$$

satisfies $\rho \neq 0$. The following statements hold:

- (i) $\nu$ is an upper bound on the fraction of outliers.
- (ii) $\nu$ is a lower bound on the fraction of SVs.
- (iii) Suppose the data were generated independently from a distribution $P(\mathbf{x})$ which does not contain discrete components. With probability 1, asymptotically, $\nu$ equals both the fraction of SVs and the fraction of outliers.

The concept of the SVDD is the following:

- Assume a data set containing $N$ objects, $\mathbf{x}_i$, and a compact description of this data is required.
- The description is given by a sphere of center $\mathbf{a}$ radius $R$, with minimum radius and which contains all (or most of) the data.
- The most outlying objects are allowed to be outside the sphere.



- The corresponding optimization problem is

$$\text{Minimize } R^2 + C\sum_{n=1}^{N}\xi_n$$
$$\text{subject to } \begin{cases} ||\mathbf{x}_n - \mathbf{a}||^2 \leq R^2 + \xi_n \\ \xi_n \geq 0 \end{cases}$$

- The solution is found by incorporating the constraints to the functional through Lagrange multipliers.

- The dual functional is

$$\begin{aligned} &\text{Minimize } -\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{z} \\ &\text{subject to } \begin{cases} \sum_{n=1}^N \alpha_n = 1 \\ \mathbf{a} = \sum_{n=1}^N \alpha_n \mathbf{x}_n \\ 0 \le \alpha_n \le 1 \end{cases} \end{aligned}$$

  where $\mathbf{z}$ is a vector containing all dot products $\mathbf{x}_n^\top \mathbf{x}_n$
- The radius $R$ can be obtained with equation $||\mathbf{x}_n - \mathbf{a}||^2 = R^2$ which will be satisfied for any $\mathbf{x}_n$ on the margin (this is, with $0 < \alpha_n < C$).

# Extensions to kernel spaces

- These two approaches are very restrictive to special distributions that satisfy:
  - The data is separable from the origin (SVND)
  - The data is contained in a *small* sphere of radius $R$ (SVDD).
- Obviously, practical cases do not fit these properties. The algorithms are described in Reproducing Kernel Hilbert Spaces where these conditions can be satisfied.
- For the case of the square exponential kernel

$$< \varphi(\mathbf{x}_n)^\top \varphi(\mathbf{x}_m) >= k(\mathbf{x}_n, \mathbf{x}_m) = exp\left(\frac{||\mathbf{x}_n - \mathbf{x}^m||^2}{2\sigma^2}\right)$$

  both algorithms are equivalent.
- The RKHS versions are to be developed in next chapter.